Chapter 7

Two Variable Linear Regression

In Chapters 3, 4, and 5 we examined the historic growth patterns for Process Control Company's sales; we recognized that this traditional time-series component analysis was inadequate for forecasting, due to the strong business cycle fluctuations in the Process Control data. Thus we now turn to regression analysis in its simplest form (two variable linear) as an extension of time-series analysis.

Regression analysis measures the numerical association between the independent explanatory variable and the dependent sales variable. The objective of this statistical approach is to forecast sales based on an equation describing the historical response of sales to an activity variable in the marketplace.

7.1 The Two Variable Linear Regression Model

When we endeavor to predict sales, Y, based on the value of the explanatory variable, X, the quantity we are seeking is expected sales, E(y), for a predetermined value X, say X_0. Hence, regression analysis is approached from the standpoint of drawing inferences from a particular set of sample (historical) observations to the population (or underlying) relationship. The population is represented by the simple linear regression model in Figure 7.1, consisting of all paired (X, Y) observations. Furthermore, this population is partitioned into subpopulations of all the Y values related with each specified X. The possible values of the explanatory variable, X, are constants, fixed in advance. Thus the sales variable, Y, is random and dependent on the value specified for X.

We generalize the two variable linear regression model under these assumptions:

1. *Linearity*. The average Y for each subpopulation is the expected Y for each X value, i.e., $\mu_{Y.X}$. These conditional expected values fall in a straight line defining the population equation Y regressed on X:

$$\mu_{Y.X} = B_1 + B_2X. \tag{7.1}$$

In this linear model, B_1 and B_2 are population parameters: B_1 is the Y-intercept (expected sales when X is zero), and B_2 is the regression line slope (the change in Y for one unit change in X). The parameter B_2 is the regression coefficient. See Figure 7.2.

2. *Normality*. All of the Y subpopulations are normally distributed.

3. *Homoscedasticity*. All of the Y subpopulations have the same variance, assuring uniform dispersion of the points about the line of regression:

$$\sigma_{Y.1}{}^2 = \sigma_{Y.2}{}^2 = \ldots = \sigma^2_{Y.X} \tag{7.2}$$

4. *Independence*. The $Y/_{\mu y.x}$ values are statistically independent of X.

Recognizing by the nature of the expected value concept that it is unreasonable for all individual Y values to fall on the line $\mu_{Y.X} = B_1 + B_2X$, we must refine the predicting equation for sales as:

$$\mu_{Y.X} = B_1 + B_2X + \epsilon_i \tag{7.3}$$

where,

ϵ_i = residual errors of each individual Y from the expected value of Y, or, $\epsilon_i = (Y_i - \mu_{Y.X})$.

Continuing, ϵ_i itself is an independent random variable, normally distributed with an expected value of zero and a constant variance for all *i* observations. Of course, in practice we do not know the population regression line. Since a straight line is defined by its intercept, B_1, and slope, B_2, our task is to approximate the population regression line by deriving estimates for B_1 and B_2. These estimates are obtained from paired sample observations so the sample regression line serves as our estimate of the population regression line, i.e., the population regression model:

$$\mu_{Y.X} = B_1 + B_2X + \epsilon_i \tag{7.4}$$

is estimated by the sample regression model:

Figure 7.1

Simple Linear Regression Model:
Normally Distributed Y Subpopulations of Specified Values for X

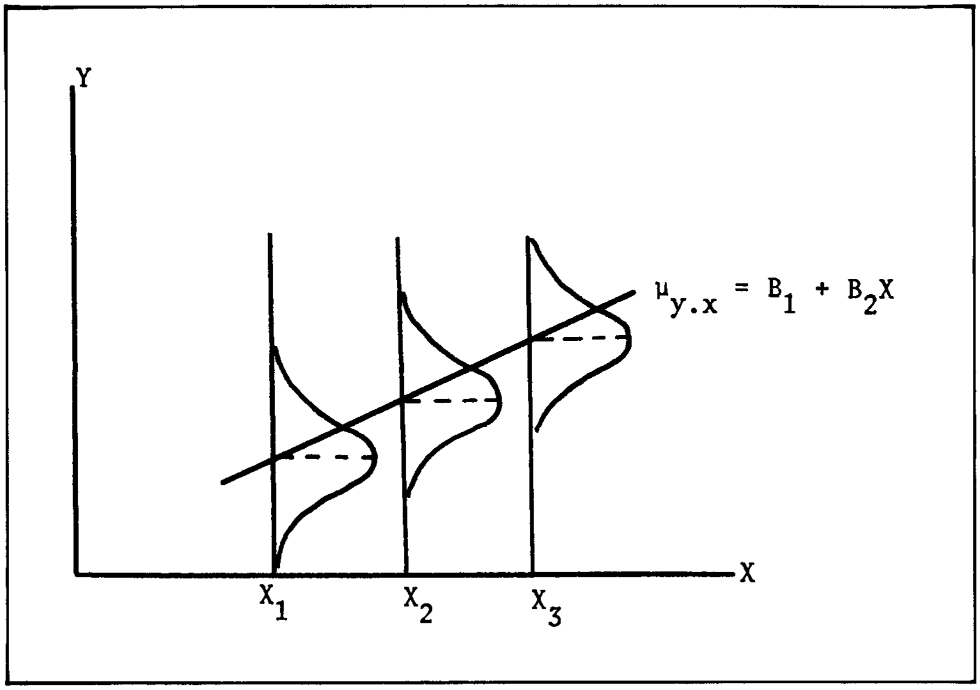
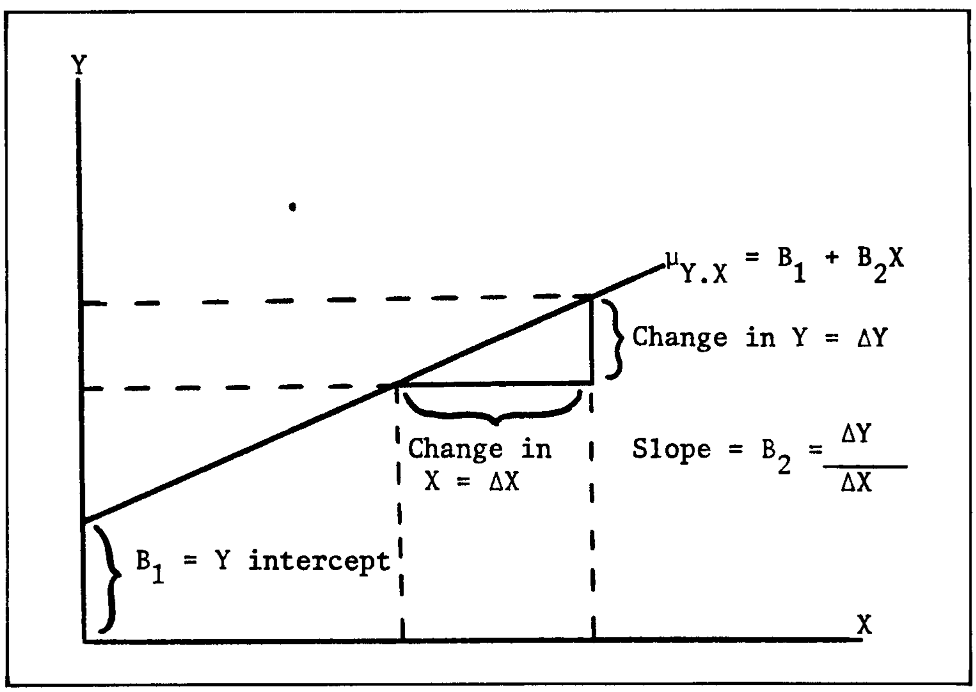


Figure 7.2

Characteristics of a Simple Linear Regression Line for Population Data



$$Y_c = b_1 + b_2X + \epsilon_i \qquad (7.5)$$

where,

Y_c = point estimate of $\mu_{Y.X}$.
b_1 = point estimate of B_1.
b_2 = point estimate of B_2.
ϵ_i = sample residual error, $(Y_i - Y_{ic})$, or simply $\epsilon_i = (Y - Y_c)$.

7.2 Case Study: Process Control Company

At this point let us reconsider Process Control Company's sales of factory machinery control devices to facilitate the discussion. We feel it would be useful to know whether there is any statistically measurable association between quarterly sales for Process Control and the hypothesized explanatory (causal) variable, New Plant and Equipment Expenditures by U.S. manufacturing durable goods industries. So that a study of these two factors can be made, we assembled the data in Table 7.1 and plotted the bivariate observations on the scatter diagram in Figure 7.3.

We can see from the scatter diagram that there is a tendency for the data points to cluster along a band extending from the lower left to the upper right and that this cluster of points suggests the existence of a positive linear relationship. Our objective is to determine a line of average relationship between the X and Y values of the points. We may approximate a straight line either by freehand methods of curve fitting or by using the objective method of least-squares. By the method of least-squares, the parameters B_1 and B_2 of the regression equation are determined such that the sum of the squared residual errors is a minimum. You will recall that residual error is defined as $\epsilon_i = (Y_i - \mu_{Y.X_i})$. Substituting the least-squares estimates, b_1 and b_2 for B_1 and B_2, and letting n be the number of sample observations, we have:

$$\sum_{i=1}^{n} \epsilon_i^2 = \text{min.} \sum_{i=1}^{n} (Y_i - b_1 - b_2X_i)^2. \qquad (7.6)$$

Hereafter we will drop the i from the Y and X notation but the meaning does not change. Resolving• this classical minimization problem results in the simultaneous solution of two normal equations:

$$\Sigma Y = nb_1 + b_2\Sigma X = \text{(Normal Equation I)} \qquad (7.7)$$

$$\Sigma XY = b_1\Sigma X + b_2\Sigma X^2 = \text{(Normal Equation II)} \qquad (7.8)$$

for the values of b_1 and b_2. The equation for b_1, the intercept, obtained by solving Normal Equation I, is written:

$$b_1 = \frac{\Sigma Y - b_2\Sigma X}{n} = \bar{Y} - b_2\bar{X}. \qquad (7.9)$$

The equation for b_2, the slope, obtained by solving Normal Equation II, is written:

$$b_2 = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}. \qquad (7.10)$$

Table 7.2 illustrates the computational procedure using the Process Control data. The resulting regression equation has been plotted on the scatter diagram in Figure 7.4. We can see already that the actual points differ from the regression line, indicating that not all the variation in sales is statistically associated with variability in the explanatory variable, New Plant and Equipment Expenditures.

Concluding this explanation of estimating the population regression line, we now enumerate the characteristics of least-squares linear regression:

1. The sum of the residual errors is zero, i.e., $\Sigma (Y - Y_c) = 0$.
2. The sum of the squares of the residual errors is minimum, i.e., $\Sigma(Y - Y_c)^2$ = minimum.
3. The computed regression line passes through the point $(\bar{X},\bar{Y})$.
4. The estimates b_1 and b_2 are unbiased, i.e., $E(b_1) = B_1$ and $E(b_2) = B_2$.

7.3 Analysis of Residual Errors: Standard Deviation of Regression

For purposes of forecasting, the usefulness of the regression line depends on the closeness of actual points to the regression line. A statistical measure showing closeness in concentration of the actual observations around the regression line is called the standard deviation of regression (standard error of estimate of Y on X). For the population this measure is:

$$\sigma^2_{Y.X} = \frac{\Sigma(Y - \mu_{Y.X})^2}{N} \qquad (7.11)$$

where,

Y = observed sales;
$\mu_{Y.X} = B_1 + B_2X$; and
N = size of the population.

Because the regression analysis is based on sample data, the approximating formula is:

$$S^2_{Y.X} = \frac{\Sigma(Y - Y_c)^2}{n - 2} \qquad (7.12)$$

when,

$Y_c = b_1 + b_2X$; and
n = number of pairs of actual observations. A more convenient calculating equation is:

$$S^2_{Y.X} = \frac{\Sigma Y^2 - b_1\Sigma Y - b_2\Sigma XY}{n - 2} \qquad (7.13)$$

Interpretation of the standard deviation of regression is similar to that of the standard deviation for any probability function. The measure provides the means to construct intervals about the regression line within which specified percentages of the actual data points may be expected to lie. For example, assuming normally dispersed observations

Table 7.1 New Plant & Equipment Expenditures, (NPEE), Manufacturing Durables, & Process Control Company Sales

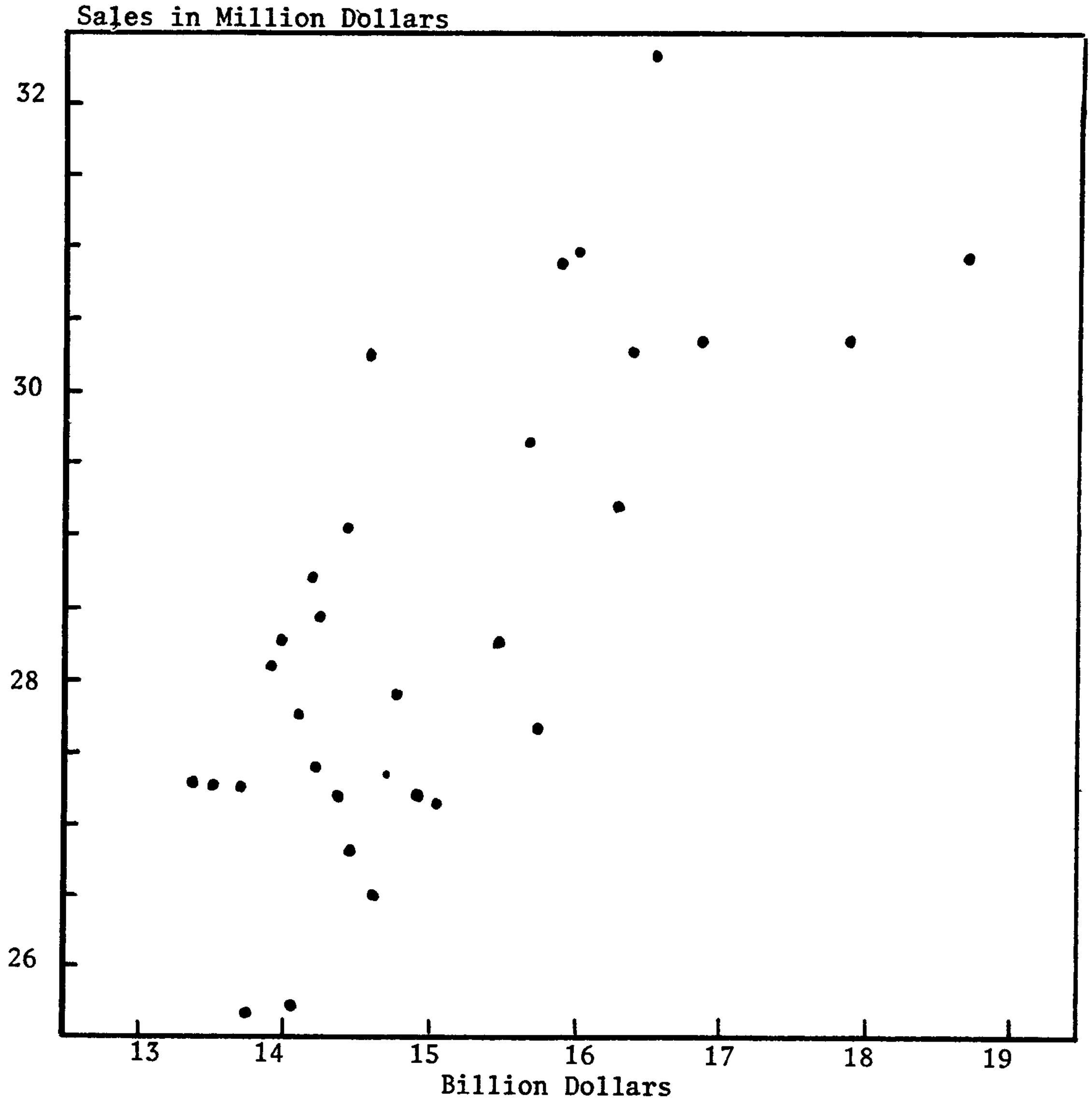
(1) Paired Observation Number	(2) Quarter & Year	(3) NPEE, Mfg. Durables, Seasonally adjusted, Annual Rate (t+2) X	(4) Quarter & Year	(5) Process Control Co. Sales, Seasonally Adjusted, (t) Y
		Billion dollars		1/10 million dollars
1	1-1966	13.28	3-1966	226
2	2-1966	13.98	4-1965	245
3	3-1966	14.18	1-1966	254
4	4-1966	14.58	2-1966	285
5	1-1967	14.46	3-1966	261
6	2-1967	14.26	4-1966	249
7	3-1967	13.92	1-1967	242
8	4-1967	13.71	2-1967	225
9	1-1968	14.11	3-1967	235
10	2-1968	13.51	4-1967	225
11	3-1968	14.47	1-1968	216
12	4-1968	14.39	2-1968	224
13	1-1969	15.47	3-1968	245
14	2-1969	15.98	4-1968	300
15	3-1969	16.53	1-1969	327
16	4-1969	15.88	2-1969	298
17	1-1970	16.40	3-1969	286
18	2-1970	16.32	4-1969	264
19	3-1970	15.74	1-1970	233
20	4-1970	14.92	2-1970	224
21	1-1971	14.21	3-1970	228
22	2-1971	14.06	4-1970	194
23	3-1971	13.76	1-1971	193
24	4-1971	14.61	2-1971	210
25	1-1972	15.06	3-1971	223
26	2-1972	14.77	4-1971	238
27	3-1972	15.67	1-1972	273
28	4-1972	16.86	2-1972	287
29	1-1973*	17.88	3-1972	287
30	2-1973*	18.70	4-1972	301

*For these two data points we used "outside" econometric forecasts for U. S. new plant and equipment expenditures, manufacturing durable goods industries.

Source: Survey of Current Business: Process Control Company.

Figure 7.3

Scatter Diagram: Process Control Company Sales and New Plant and Equipment Expenditures, Mfg. Dur.



New Plant and Equipment Expenditures, Mfg. Dur.

Table 7.2 Calculations Required for Determining Regression Line Constants

(1) NPEE, Mfg. Dur., Seasonally adj. Annual Rate (t+2) X	(2) Process Control Company Sales, Seasonally adj. (t) Y	(3) (1) x (2) XY	(4) $(2)^2$ X^2	(5) $(3)^2$ Y^2
Billion dollars	100,000 Dollars			
13.28	226	3001.28	176.36	51076
13.98	245	3425.10	195.44	60025
14.18	254	3601.72	201.07	64516
14.58	285	4155.30	212.58	81225
14.46	261	3774.06	209.09	68121
14.26	249	3550.74	203.35	62001
13.92	242	3368.64	193.77	58564
13.71	225	3084.75	187.96	50625
14.11	235	3315.85	199.09	55225
13.51	225	3039.75	182.52	50625
14.47	216	3125.52	209.38	46656
14.39	224	3223.36	207.07	50176
15.47	245	3790.15	239.32	60025
15.98	300	4794.00	255.36	90000
16.53	327	5405.31	273.24	106929
15.88	298	4732.24	252.17	88804
16.40	286	4690.40	268.96	81796
16.32	264	4308.48	266.34	69696
15.74	233	3667.42	247.75	54289
14.92	224	3342.08	222.61	50176
14.21	228	3239.88	201.92	51984
14.06	194	2727.64	197.68	37636
13.76	193	2655.68	189.34	37249
14.61	210	3068.10	213.45	44100
15.06	223	3358.38	226.80	49729
14.77	238	3515.26	218.15	56644
15.67	273	4277.91	245.55	74529
16.86	287	4838.82	284.26	82369
17.88	287	5131.56	319.69	82369
18.70	301	5628.70	349.69	90601
451.67	7498	113,838.08	6,849.96	1,853,760
Σ X	Σ Y			

$\overline{X} = 15.06$ $\overline{Y} = 249.93$

Continued on next page

$$b_2 = \frac{n \Sigma X Y - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{30\ (113{,}838.08) - (451.67)(7498)}{30\ (6849.96) - (451.67)^2}$$

$= \$19.095 \times 10^5$ per \$1 billion in equipment expenditures

$$b_1 = \overline{Y} - b_2 \overline{X}$$

$$= 249.93 - (19.095)(15.06)$$

$$= \$ - 37.551\ (\times 10^5)$$

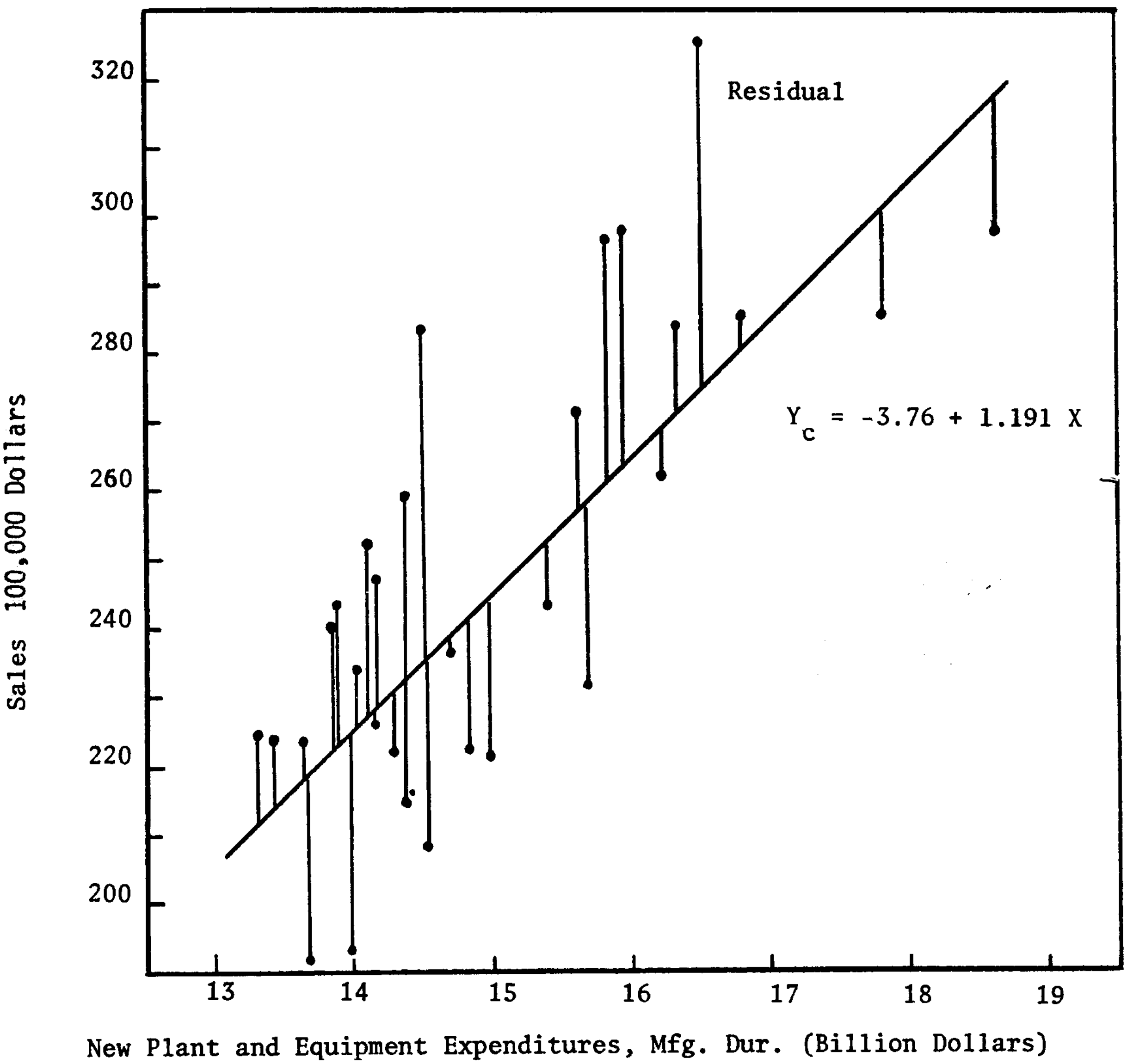
$$Y_c = b_1 + b_2 X$$

$Y_c = - 37.551 + 19.095\ X$ (predicted sales in 10^5 dollars,
based on equipment expenditures
in billion dollars)

Source: Survey of Current Business and Process Control Company

Figure 7.4 Process Control Sales and NPEE

Scatter Diagram, Regression Line, and Residuals



about the regression line, two bands parallel to the Y_c line, represented by $Y_c \pm S_{Y.X}$, include 68.27 percent of the data points; $Y_c \pm 2S_{Y.X}$ includes 95.45 percent of the items; and $Y_c \pm 3S_{Y.X}$ includes 99.73 percent.[1]

Using Equation 7.13 and numbers from Table 7.2, the computation required to determine standard deviation of regression in our study of Process Control sales and new plant and equipment expenditures is:

$$S_{Y.X} = \frac{(1{,}853{,}760) - (-37.551)(7498) - (19.1)(113{,}838)}{30 - 2}$$

$$= 23.609.$$

In Figure 7.5 we show the band that will include about 68 percent of the data points. This is an approximation, and we will give more exact formulas later under the discussion of confidence intervals.

7.4 Inferences From the Regression Line Slope

The population coefficient of regression, B_2, is interpreted as the average change in sales, Y, for a unit change in the explanatory variable, X. It is, therefore, an important measure of association between these variables.

Frequently, the first inference with which we deal in a regression study involves whether the value of B_2 is significant. If $B_2 = 0$, then the population regression line is horizontal, implying no relationship between X and Y; i.e., changes in X have no influence on the values assumed by Y.

Student's *t* distribution is the basis for measuring the statistical significance of B_2. The set of hypotheses are stated, as follows: Null Hypothesis–$B_2 = 0$, and Alternative Hypothesis–$B_2 \neq 0$. Using a two-sided test, and an appropriate level of significance and degrees of freedom (n – 2), the critical value of *t* is determined from Appendix B. Then the *t* value from the sample data is computed using:

$$t = \frac{b_2 - B_2}{S_{b_2}} \qquad (7.14)$$

where the standard error of the sampling distribution of b_2 is estimated by

$$S_b = \frac{S_{y.x}}{\sqrt{\Sigma(X - \bar{X})^2}} = \frac{S_{y.x}}{\sqrt{\Sigma X^2 - n\bar{X}^2}} \qquad (7.15)$$

Testing B_2 in the Process Control study, at a 0.05 significance level and 30 – 2 = 28 degrees of freedom, the critical *t* value equals 2.048. The rejection region is $t < -2.048$ and $t > +2.048$. The *t* value calculated from the sample data is:

$$t = \frac{19.095 - 0}{\dfrac{\sqrt{23.609}}{6849.96 - 30(15.056)^2}} = \frac{19.095}{3.345} = 5.71$$

Because $t = 5.71 > t_{.025;28} = 2.048$, the null hypothesis $B_2 = 0$ can be rejected. Therefore the slope b_2 is significant and is statistical evidence of a relationship between sales and new plant and equipment expenditures.

In addition, we can establish an interval estimate for B_2 as:

$$\text{Confidence limits for } B_2 = b_2 \pm tS_{b_2} \qquad (7.16)$$

For a 95 percent confidence interval the confidence limits are:

$19.095 \pm (2.048)(3.345)$,
12.244 to 25.946.

We conclude that sales for Process Control Company increase from between $\$12.244(\text{x}10^5)$ to $\$25.946(\text{x}10^5)$ for each billion dollar increase in new plant and equipment expenditures.

7.5 Measuring Quality of the Relationship

Although we have established that the slope, b_2, of the regression line is statistically significant, this information gives us no insight into the "degree" to which the sales variable is linearly related to new plant and equipment expenditures. The objective of this section is to introduce methodology that shows how closely the Y and X variables are associated in the simple linear regression model.

One measure of the usefulness of the Y_c regression line is provided by comparing the standard deviation of regression with the standard deviation of Y. The standard deviation of Y measures the dispersion of Y's around the horizontal $\bar{Y}$ line before the explanatory variable X is introduced. Considered as a measure of total variability in the Y's, it is estimated from sample data by:

$$S_Y = \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{n - 1}} = \sqrt{\frac{\Sigma Y^2 - (\Sigma Y)^2/n}{n - 1}} \qquad (7.17)$$

Substitution of sales values from Process Control's data given in Table 7.2 yields:

$$S_Y = \sqrt{\frac{1{,}853{,}760 - (7498)^2/30}{30 - 1}} = 33.548$$

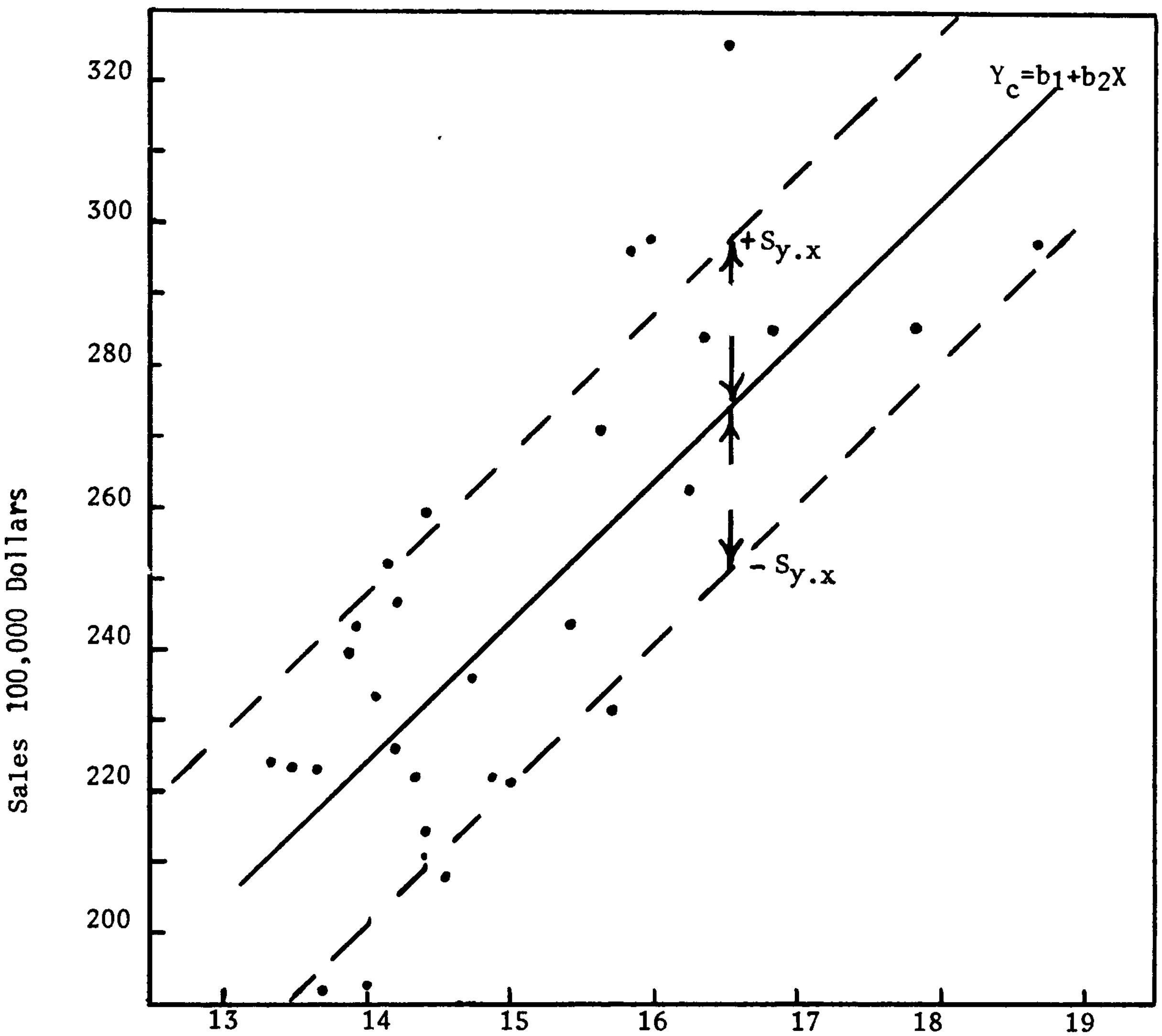
The standard deviation of regression $S_{Y.X}$ was calculated previously as $\$23.609(\text{x}10^5)$ or slightly more than 70 percent of the standard deviation of Y, $\$33.548(\text{x}10^5)$. These figures show that when new plant and equipment expenditures are employed to predict sales, the dispersion in Y decreases nearly 30 percent.

We may visualize the closeness of the relationship between the two variables by comparing the ranges encompassed by both the standard deviation of regression and the standard deviation of Y in Figure 7.6. Since the vertical distance representing $S_{Y.X}$ is smaller than that representing S_Y, the measure of total dispersion in the Y's, we say there exists a good association; i.e., the regression line, Y_c, has helped explain much of the scatter in the Y's. Poor association, in contrast, would be represented by a vertical distance for $S_{Y.X}$ nearly as large as that for S_Y.

Hughes and Grawoig provide the following summary of the general idea of this type of scatter diagram analysis:

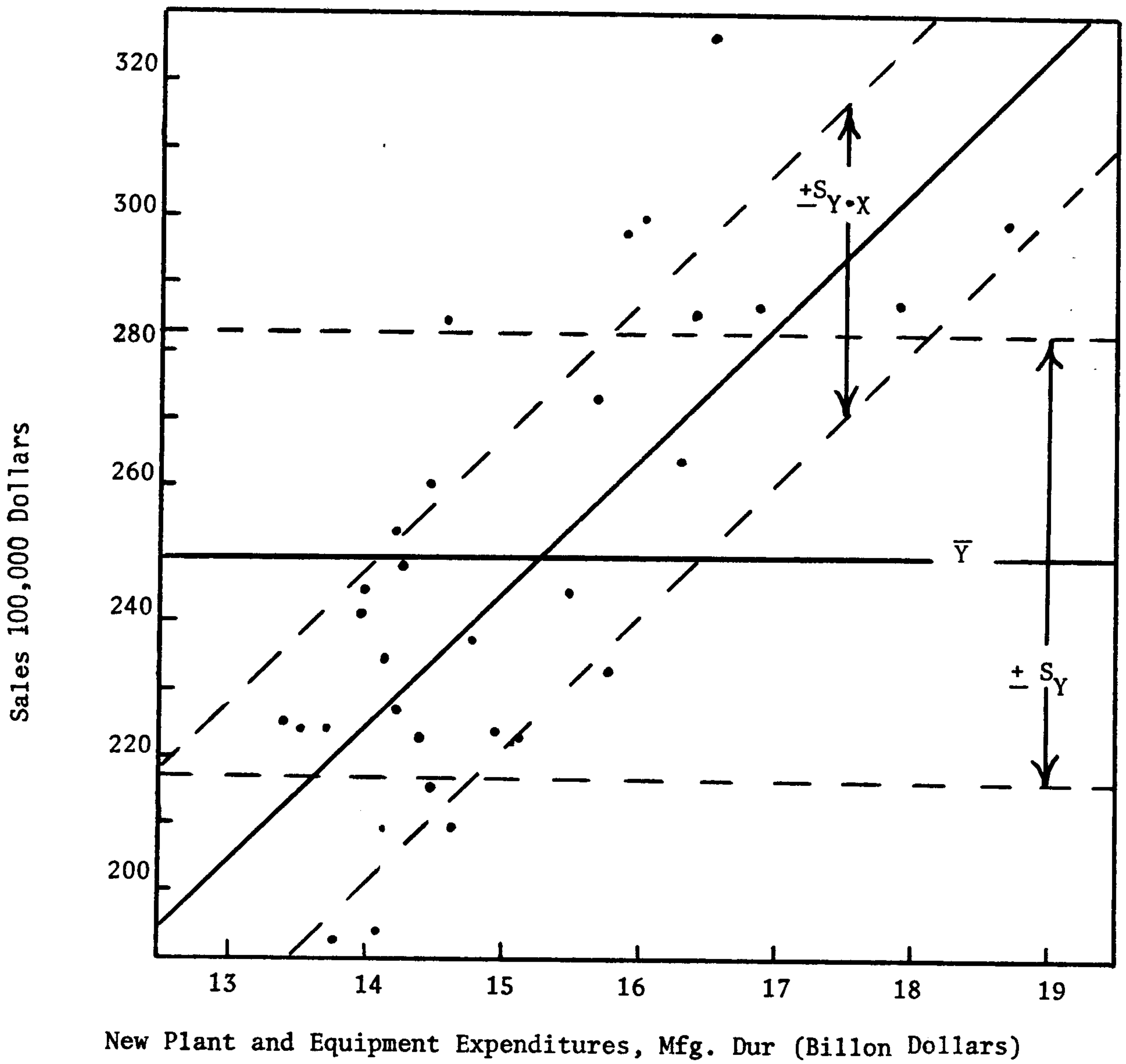
Figure 7.5

Scatter Diagram and Standard Error of Regression; Process Control Company



New Plant and Equipment Expenditures, Mfg. Dur. (Billion Dollars)

Figure 7.6
Standard Deviation of Regression, $S_{Y.X}$, Compared to Standard Deviation of S_Y : Good Association



The standard deviation of Y, S_Y, measures the scatter of the Y values around their mean; the standard deviation of regression, $S_{Y.X}$, measures the scatter of the Y values around the line of average relationship. If the series of values represented by the regression line is a better description of the relationship between the variables than is the single value that the Y provides, the dispersion of the values around the line will be less than around the mean. The less accurately the estimating line describes the relationship between the variables, the greater will be the extent of the dispersion around the line. If the line perfectly depicts the relationship, all the actual values will coincide exactly with the estimates and there will be no dispersion at all around the line. If no correlation exists, the dispersion around the estimating line will be as great as around the mean of the Y values.[2]

7.6 Coefficients of Determination of Correlation

We can analyze the dispersion of the Y values to evolve a concept of goodness-of-fit of the regression line to observed data. Total variation in Y may be *partitioned* into dispersion which is *explained* and *unexplained* by the regression line. From Figure 7.7 we establish the relationship:

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - Y_c)^2 + \Sigma(Y_c - \bar{Y})^2 \tag{7.18}$$

which is verbally interpreted as

$$\begin{pmatrix}\text{Total}\\ \text{sum of}\\ \text{squares}\end{pmatrix} = \begin{pmatrix}\text{Unexplained}\\ \text{sum of}\\ \text{squares}\end{pmatrix} + \begin{pmatrix}\text{Explained}\\ \text{sum of}\\ \text{squares}\end{pmatrix} \tag{7.19}$$

The $\Sigma(Y - Y_c)^2$ is termed unexplained since it is the part of the original dispersion that remains after fitting the regression line through the data. On the other hand, $\Sigma(Y_c - \bar{Y})^2$ is called explained since it is the part of the original dispersion that is eliminated when the regression line has been fitted.

Because we are interested in the relationship between $\Sigma(Y_c - \bar{Y})^2$ and $\Sigma(Y - \bar{Y})^2$, we define the ratio

$$\tilde{r}^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = \frac{\Sigma(Y_c - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} \tag{7.20}$$

$$= 1 - \frac{\Sigma(Y - Y_c)^2}{\Sigma(Y - \bar{Y})^2}$$

This ratio, based on sample data, is called the estimated coefficient of determination and is interpreted as the percent or fraction of total variation in Y explained by X in the regression model. So, the closer $\tilde{r}^2$ is to ±1, the closer the data points tend to fall about the regression line.

When $\tilde{r}^2$ is calculated using Equation 7.20, it is positively biased, especially for a small sample of observations. Adjusting for the bias, the sum of squares can be divided by their respective degrees of freedom. Consequently,

$$r^2 = 1 - \frac{\Sigma(Y - Y_c)^2/(n-2)}{\Sigma(Y - \bar{Y})^2/(n-1)} = 1 - \frac{S^2_{Y.X}}{S^2_Y} \tag{7.21}$$

Notice that as sample size gets large, (n – 2) and (n – 1) effectively cancel one another, causing $\tilde{r}^2$ to equal r^2.

Using numbers from Table 7.3 results in:

$$r^2 = 1 - (557.4/1164.3)$$
$$= 0.538$$

This r^2 indicates that about 54 percent of the variation in Process Control sales is linearly related with variation in new plant and equipment expenditures as described by the regression model:

$$Y_c = -37.544 + 19.095X \tag{7.22}$$

Taking the square root of the coefficient of determination, r^2, we obtain the coefficient of correlation, r. A computational form for $\tilde{r}$ is:

$$\tilde{r} = \sqrt{\frac{n\Sigma XY - \Sigma X\Sigma Y}{[n\Sigma X^2 - (\Sigma X)^2]\ [n\Sigma Y^2 - (\Sigma Y)^2]}} \tag{7.23}$$

The sign attached to *r* is the sign of b_2 in the regression equation. Hence, *r* is positive when the regression line has an upward slope. Correspondingly, it is negative when the regression line has a downward slope. When r = 0, we say there is no correlation (association or relationship) between X and Y.

The possible values for the coefficient of correlation range from –1 to +1. If all data points fall on the regression line there is perfect correlation, and r = +1 or –1. However, when the scatter of points is such that the least-square's line is horizontal (coincident with $\bar{Y}$), then r is zero and there is no correlation.

7.7 A Significance Test for *r*

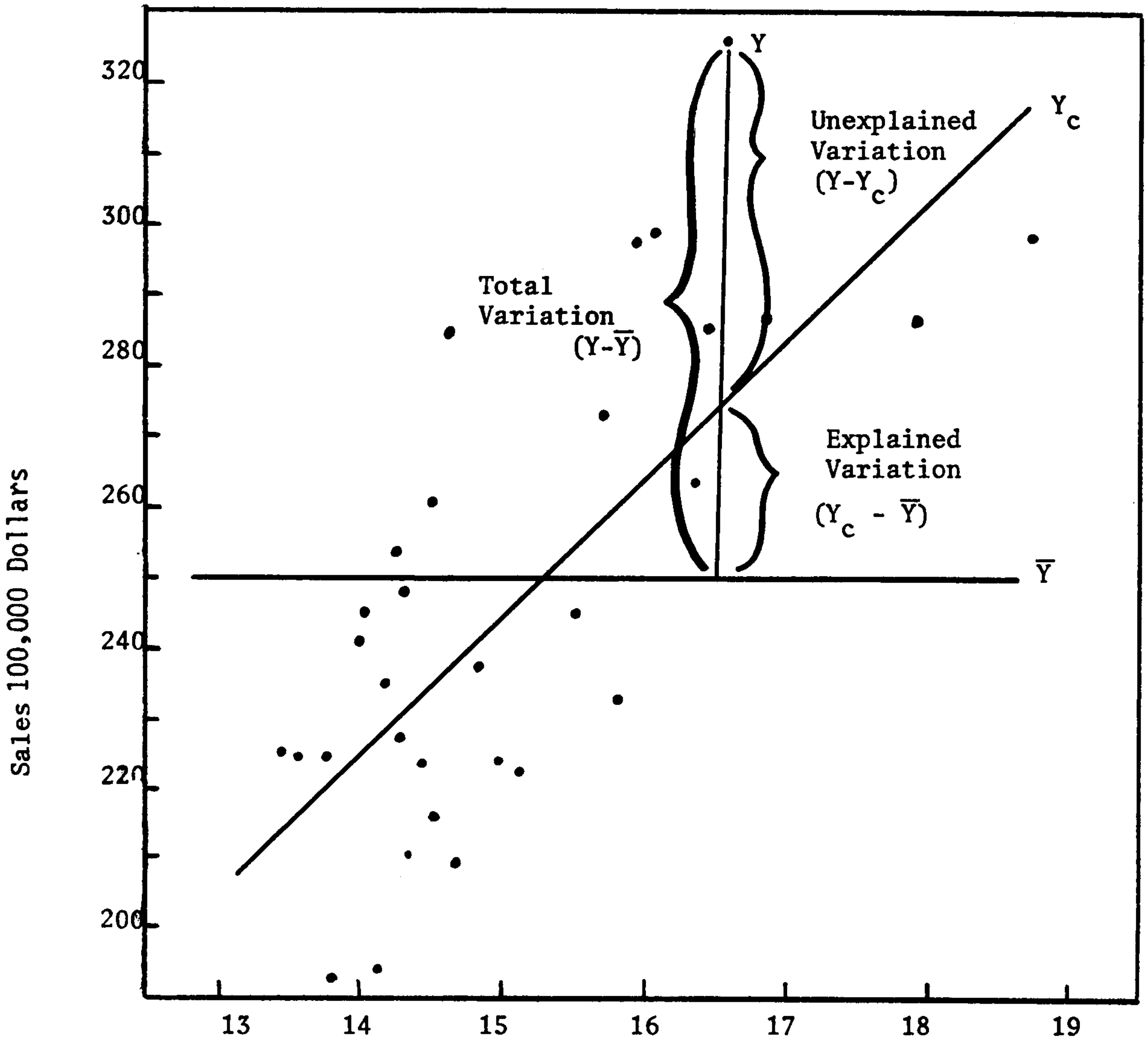
The coefficient of correlation for the Process Control study is $r = 0.733$. Because this *r* value is computed from sample data, we must test if it is statistically significant. The hypothesis of no association between the X and Y variables (Null: $\rho = 0$, where ρ is the population coefficient of correlation) can be accepted or rejected based on the F ratio in Table 7.3. The statistic is calculated by

$$F = \frac{\text{Explained Variance}}{\text{Unexplained Variance}} = \frac{\Sigma(Y_c - \bar{Y})^2/1}{\Sigma(Y - Y_c)^2/(n-2)} \tag{7.24}$$

We are interested in determining whether or not the calculated value of F at a specified significance level is larger than the critical value obtained from the F table in Appendix C. Since the test statistic, F = 32.575, is greater than the critical value, F = 4.2 (numerator degrees of freedom = 1; denominator degrees of freedom = 28; and significance level = 0.05), the null hypothesis of no correlation is rejected, and we accept Alternative: $\rho \neq 0$; i.e., there is a statistically significant linear correlation between Process Control Company sales and new plant and

Figure 7.7

Partitioning Total Variation



New Plant and Equipment Expenditures, Mfg. Dur. (Billion Dollars) equipment expenditures. The regression line, therefore, better describes the data and provides a better basis for predicting sales than the historic average level of sales.

7.8 Autocorrelation

One critical assumption of the linear regression model is independence of the successive residual error terms, ϵ_i. Any deficiency in this assumption defines the undesirable property of autocorrelation. The existence of autocorrelated residual errors indicates that some factor(s) present in the sales variable has not been explained by the regression model.

In order to test for the presence of autocorrelation the Durbin-Watson statistic is calculated by:

$$d = \frac{\sum_{i=2}^{n} (\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n} \epsilon_i^2} \qquad (7.25)$$

where,

ϵ_i = unexplained residual error from the regression model for observation i with n observations.

Hence, the Durbin-Watson statistic equals the sum of the squares of what is called "first differences" of the residual errors divided by the sum of squares of the residual errors.

After computing the d statistic, one of the tables provided in Appendix D is selected depending on the desired significance level. These tables are used to determine critical values for a two-tailed hypothesis test where n is the number of paired data points in the regression analysis and k' is the number of explanatory variables in the equation.

Calculating d enables a comparison of this statistic with appropriate critical values for d_1 and d_u (or $4-d_u$ and $4-d_L$) from Appendix D. Referring to the generalized bottom scale of Figure 7.7a and reading left to right:

Testing for positive autocorrelation (when $d < 2$):

If $0 < d < d_L$, significant positive autocorrelation exists.

If $d_L < d < d_u$, test results are indecisive; i.e., no decision.

If $d_u < d < 2$, no significant positive autocorrelation exists.

Testing for negative autocorrelation ($d > 2$):

If $2 < d < 4-du$, no significant negative autocorrelation exists.

If $4-du < d < 4-dL$, test results are indecisive.

If $4-d_L < d < 4$, significant negative autocorrelation exists.

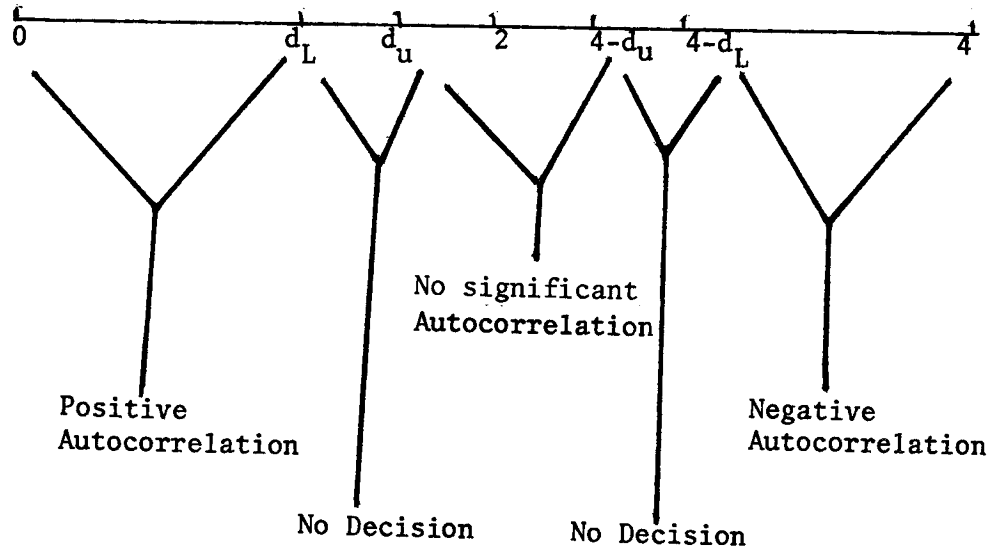
Calculations for d for the Process Control analysis are shown in Table 7.4. With a 0.05 level of significance and $n = 30$ and $k' = 1$, the critical values $d_L = 1.25$ and $d_u = 1.38$ are read from Appendix D. The conclusion of statistically significant positive autocorrelation is thus established for our linear regression model since $d = 0.552 < d_L = 1.25$. This adverse result implies the necessity for further analysis, possibly either by (1) introducing other explanatory variables (see Chapter 8); (2) constructing an improved

Table 7.3

Partitioning Total Sum of Squares into Explained and Unexplained Sum of Squares

(1) Type of Variation		(2) Sum of Squares	(3) Degrees of Freedom	(2)÷(3)=(4) Variance or Mean Square	(5) F Ratio (Explained Variance ÷ Unexplained Variance)
Explained	$\Sigma(Y_c-\bar{Y})^2$	18156.9	1	18156.9	32.575
Unexplained	$\Sigma(Y-Y_c)^2$	15607.1	28	557.4	
Total	$\Sigma(Y-\bar{Y})^2$	33764.0	29	1164.3	

Figure 7.7a: Autocorrelation Test Scale for Process Control Analysis



functional form (Chapter 8); (3) transforming variables (Chapter 9); or (4) some combination of these.

7.9 Interval Estimates of Prediction

Before continuing with needed refinements in the present regression model, for instructional purposes we consider methodology for determining predictions based on the least-squares equation. We consider first the case of estimating the *mean* or average sales, Y, for a given value of the explanatory variable, X. Appropriate interval estimates for expected sales can be written:

$$\text{Confidence Limits for } \mu_{Y.X} = Y_c \pm t \cdot S_{Y.X}\sqrt{\frac{1}{n}+\frac{(X_e-\overline{X})^2}{\Sigma X^2 - n\overline{X}^2}} \qquad (7.26)$$

where,

X_e = the value of the explanatory variable used as an input estimator.

Applying this theory to predicting Process Control sales for new plant and equipment expenditure, X_e = \$18 billion, the predicted sales would be $Y_c = 37.551 + 19.095(18) =$ \$306.155(x10^5). To find 95 percent confidence limits, we obtain

$$306.155 \pm (2.048)(23.609)\sqrt{\frac{1}{30}+\frac{(18-15.056)^2}{6849.96-30(15.056)^2}}$$

or 284.137 to 328.172.

In other words, presuming validity for the regression model, we assert with a probability of 0.95 that average sales, when new plant and equipment expenditures are \$18 billion, will be contained in the interval from \$284.137(x10^5) to \$328.172(x10^5).

In most cases, where the prediction of an *individual* Y value on a given X is desired, we have

$$\text{Confidence Limits for } Y = Y_c \pm t \cdot S_{Y.X}\sqrt{1+\frac{1}{n}+\frac{(X_e-\overline{X})^2}{\Sigma X^2 - n\overline{X}^2}} \qquad (7.27)$$

This equation differs from Equation 7.26 only in the first term under the radical sign.

To further illustrate, consider Process Control sales for a *particular* time period, given \$18 billion in new plant and equipment expenditures. What would be predicted sales? Notice we are not asking about *average* sales for Process Control Company, rather, we now are inquiring about the sales for an *individual* time period in the future. Hence, the confidence interval may be stated:

$$306.155 \pm (2.048)(23.609)\sqrt{1+\frac{1}{30}+\frac{(18-15.056)^2}{6849.96-30(15.056)^2}}$$

= 253.026 to 359.283.

On a given X, the predicted confidence interval for individual sales is wider than the confidence interval for average sales. This is always the case since the wider interval

Table 7.4 Calculations Required for Determining the Durbin-Watson Statistic for Process Control Company Sales Forecasts.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Y	Y_c	$(Y-Y_c) = \varepsilon_i$	ε_{i-1}	$\varepsilon_i-\varepsilon_{i-1}$	$(\varepsilon_i-\varepsilon_{i-1})^2$	ε_i^2
226	216.027	-9.973	·			99.461
245	229.394	-15.606	- 9.973	- 5.633	31.731	243.547
254	233.213	-20.787	-15.606	- 5.181	26.843	432.099
285	240.851	-44.149	-20.787	-23.362	545.783	1949.134
261	238.559	-22.441	-44.149	21.708	471.237	503.598
249	234.740	-14.260	-22.441	8.181	66.929	203.348
242	228.248	-13.752	-14.260	0.508	0.258	189.118
225	224.238	- 0.762	-13.752	12.990	168.740	0.581
235	231.876	- 3.124	- 0.762	- 2.362	5.579	9.759
225	220.419	- 4.581	- 3.124	- 1.457	2.123	20.986
216	238.750	22.750	- 4.581	27.331	746.984	517.562
224	237.223	13.223	22.750	- 9.527	90.764	174.848
245	257.845	12.845	13.223	- 0.378	0.143	164.994
300	267.583	-32.417	12.845	-45.262	2048.647	1050.862
327	278.085	-48.915	-32.417	-16.498	272.184	2392.677
298	265.674	-32.326	-48.915	16.589	275.195	1044.970
286	275.603	-10.397	-32.326	21.929	480.881	108.098
264	274.075	10.075	-10.397	20.472	419.103	101.506
233	263.000	30.000	10.075	19.925	397.006	900.000
224	247.343	23.343	30.000	- 6.657	44.316	544.896
228	233.786	5.876	23.343	-17.557	308.248	33.478
194	230.921	36.921	5.786	31.135	969.388	1363.160
193	225.193	32.193	36.921	- 4.728	22.354	1036.389
210	241.423	31.423	32.193	- 0.770	0.593	987.405
223	250.016	27.016	31.423	- 4.407	19.422	729.864
238	244.478	6.478	27.016	-20.538	421.809	41.964
273	261.664	-11.336	6.478	-17.814	317.339	128.505
287	284.386	- 2.614	-11.336	8.722	76.073	6.833
297	303.863	16.863	- 2.614	19.477	379.353	284.361
301	310.521	18.521	16.863	1.658	2.749	343.027
					8,611.774	15,607.030

$$d = \frac{8611.774}{15607.030}$$

$$d = 0.552$$

Source: Table 7.1 and Y_c calculated from Equation 7.22

Figure 7.8

Confidence Limits for Average Sales, $M_{Y.X}$, and Individual Sales, Y

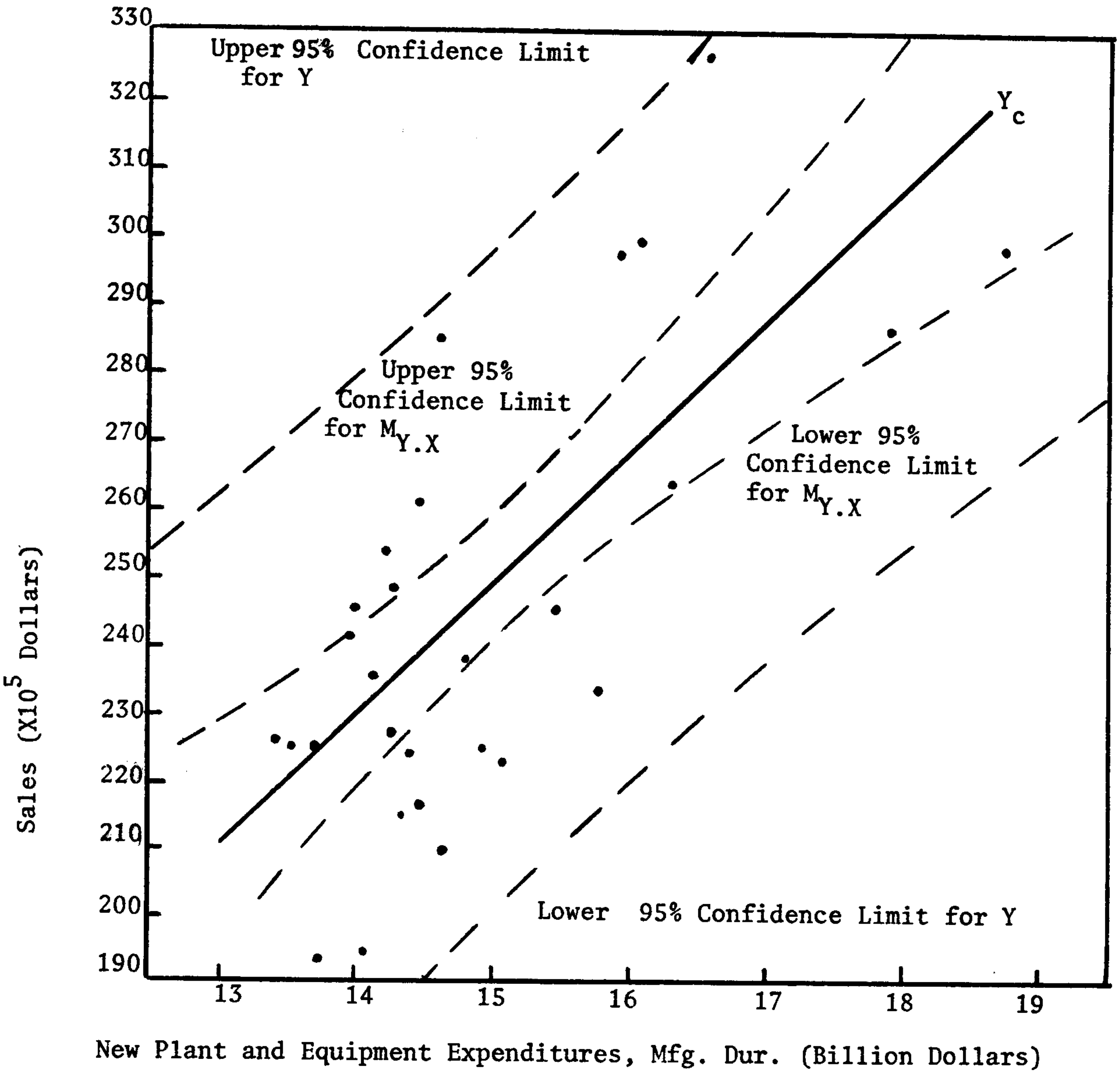


Table 7.5

Simple Regression* Forecasts: Process Control Company Sales

Based on New Plant and Equipment Expenditures, Mfg. Dur.

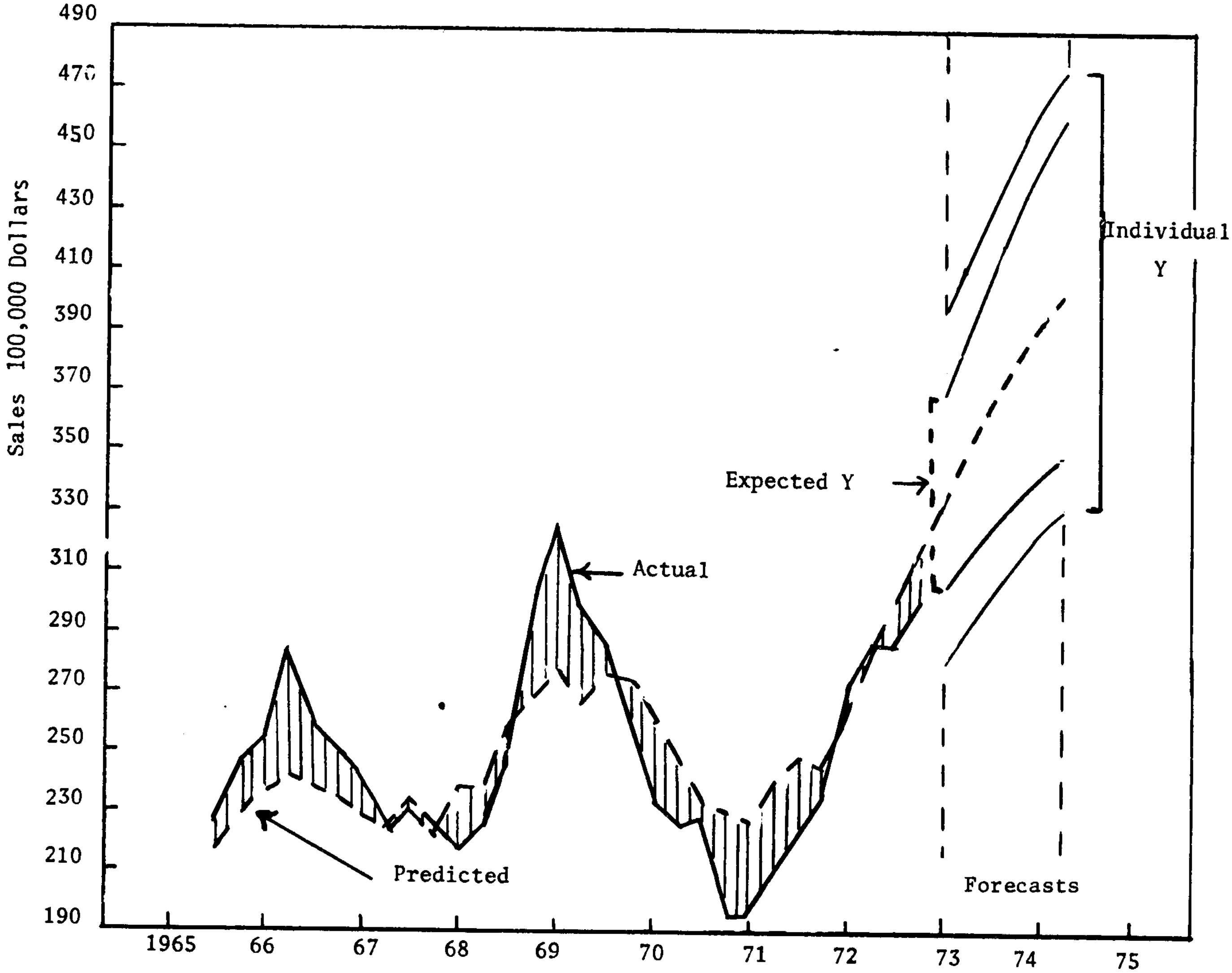
Quarter & Year	Projected** Input Values for NPEE, Mfg. Dur.	Process Control Company Sales Forecasts, 95% Confidence Limits (1/10 million dollars)			
		Expected Sales, $\mu_{Y \cdot X}$		Individual Sales, Y	
	bil. dol.	Upper	Lower	Upper	Lower
1-1973	19.7	371.5	305.5	397.1	280.0
2-1973	20.5	393.2	316.1	416.5	292.8
3-1973	21.3	413.8	325.9	435.2	304.5
4-1973	22.1	433.0	335.2	452.9	315.3
1-1974	22.8	451.0	343.8	469.6	325.2
2-1974	23.3	463.9	349.9	481.6	332.1

* Y = - 37.551 + 19.095X
** Obtained from "outside" econometric model.

Figure 7.9

Process Control Company:

Actual and Predicted Sales for 1965-1972, Sales Forecasts for Six Quarters of 1973-1974, and Confidence Intervals for Six Quarters of Forecasts of Expected Sales and Individual Sales



reflects the past variability of individual quarters, rather than the past variability of the average of many quarters. Thus individual estimates of Y are always less precise than estimates for means of Y, all based on a particular value of X. For purposes of visual comparison in Figure 7.8, we have constructed 95 percent confidence limits for estimating average and individual sales for Process Control Company.

Usually in preparation of a short-range forecast, at least five quarters into the future are desired. To illustrate, we prepared forecasts (recorded in Table 7.5) for 1973 and half of 1974 by quarters, based on simple linear regression. We note that the 95 percent confidence limits for forecasts of sales variable, Y (shown graphically in Figure 7.9), are based exclusively on measures of past statistical error and, therefore, do not include either possible errors in the future input values of the explanatory variable X or possible changes in the business or economic structural relationships that underlie use of b_1 and b_2 calculated from past data.

7.10 Words of Caution: Causality and Statistics

Realistic forecasts which contribute greatly to both individual company success and to the stability of the entire economy are the results of applying sound business experience and judgment to relevant and timely statistical analyses. We emphasize that regression analysis does not of itself prove economic causality; it only measures the degree of mathematical association between the recorded data for sales and the explanatory variable(s). Hence a regression equation should not be used as a predicting device unless there is a rational causal relationship underlying the predicting equation.

We also emphasize that a regression model is an approximation that is most useful over the range for which past observations are available. Extrapolation for predicting, therefore, is hazardous since the association among variables is more likely to change for predictions outside the range of historical data than inside.

As a final word of caution, we point out that the use of an established statistical regression equation for forecasting assumes no change in past relationships over future time. Moreover, the utilization of an equation presumes no unusual events (e.g., economic controls, scarcity of input resources, or the like) which would tend to reduce the forecasting accuracy of the regression model. Hence, even though a forecast may be comfortably within the range of past observations, the forecaster must be constantly perceptive of the limitations of the underlying static quality of his model.

Footnotes

1. For small numbers of data points, however ($n < 30$), Student's *t* is the theoretically correct sampling distribution rather than the standard normal.

2. Ann Hughes and Dennis Grawoig, *Statistics: A Foundation for Analysis* (Reading, Massachusetts, Addison-Wesley Publishing Company, 1971), pp. 342-343.

Bibliography

Benton, William K. *Forecasting for Management.* Reading, Massachusetts: Addison-Wesley Publishing Company, 1972, ch. 3.

Chisholm, Roger K. and Gilbert R. Whitaker, Jr. *Forecasting Methods.* Homewood, Illinois: Richard D. Irwin, Inc., 1971, ch. 7.

Chou, Ya-lun. *Statistical Analysis with Business & Economic Applications.* New York: Holt, Rinehart and Winston, Inc., 1969, ch. 17 and 19.

Clark, Charles T. and Lawrence L. Schkade. *Statistical Methods for Business Decisions.* Dallas, Texas: South-Western Publishing Company, 1969, ch. 16.

Dauten, Carl A. and Lloyd M. Valentine. *Business Cycles and Forecasting.* Dallas, Texas: South-Western Publishing Company, 1968, ch. 10.

Enrick, Norbert Lloyd. *Market and Sales Forecasting.* San Francisco: Chandler Publishing Company, 1969, ch. 7.

Hughes, Ann and Dennis Grawoig. *Statistics: A Foundation for Analysis.* Reading, Massachusetts: Addison-Wesley Publishing Company, 1971, ch. 14 and 15.

Mason, Robert D. *Statistical Techniques in Business and Economics*, rev. ed. Homewood, Illinois: Richard D. Irwin, Inc., 1970, ch. 16.

Neter, John; William Wasserman; and G.A. Whitmore. *Fundamental Statistics for Business and Economics.* Boston: Allyn and Bacon, Inc., 1973, ch. 22 and 23.

Peters, William S. and George W. Summers. *Statistical Analysis for Business Decisions.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1969, ch. 14 and 16.

Salzman, Lawrence. *Computerized Economic Analysis.* New York: McGraw-Hill, 1968, ch. 4.

Spurr, William A. and Charles P. Bonini. *Statistical Analysis for Business Decisions.* Homewood, Illinois: Richard D. Irwin, Inc., 1967, ch. 22.

Stockton, John R. and Charles T. Clark. *Introduction to Business and Economic Statistics.* Dallas, Texas: South-Western Publishing Company, 1971, ch. 11.

Thompson, Gerald E. *Statistics for Decisions, An Elementary Introduction.* Boston: Little, Brown and Company, 1972, ch. 19.

Wonnacott, Thomas H. and Ronald J. Wonnacott. *Introductory Statistics for Business and Economics.* New York: John Wiley and Sons, Inc., 1972, ch. 11, 12, and 14.